Joonas Tuominen

Merger of Digital Journalistic Archives

Metropolia University of Applied Sciences

Bachelor of Engineering

Degree Programme in Media Engineering

Thesis

7 November 2017

Metropolia

Author	Joonas Tuominen
Title	Merger of digital journalistic archives
Number of Pages	35 pages + 2 appendices
Date	7 November 2017
Degree	Bachelor of Engineering
Degree Programme	Media Engineering
Instructors	Saara Dahlbacka, System Manager Aarne Klemetti, Researching Lecturer

The object of this thesis is to document the project of designing and carrying out the transfer and merger of the contents of a digital journalistic archive with another digital archive. A proper archive of published documents provides the client company a reliable source to find information when creating new content. The project included designing parts of a user interface, the selection and connection of metadata fields, creating a metadata recognizing algorithm, and carrying out the physical task of exporting the contents of an archive. The thesis investigates principles of archiving to apply them in practice.

The goal of the project was to improve the user experience of accessing the contents of the transferable archive and to modernize its outlook while maintaining the integrity of the stored metadata. Merging the archives and improving the end product required brainstorming, designing, and understanding of databases and programming.

The project was arranged with individuals joining from the client organization to help in the specialized points of interest. Working on the project was generally done through Skype in co-operation with a system specialist. The project required usage of programs that included both the old and new file frameworks (Doris and Trip respectively), DbEdit and UgEdit for adjusting the old archive to make it appropriate for the merger, and DbTest for accessing databases, as well as the use of Microsoft Office programs.

The final implementation is a functional archive which retains the original metadata of documents while giving users a better user experience, better presentation of the contents of the archive and enhanced search functions. The principles this thesis recommends can be applied to other mergers of digital journalistic archives to produce an equally functional archive.

Keywords	archive, digital archive, metadata, transfer

Metropolia

Tekijä Otsikko Sivumäärä Aika	Joonas Tuominen Digitaalisten journalististen arkistojen yhdistäminen 35 sivua + 2 liitettä 7.11.2017
Tutkinto	Bachelor of Engineering
Koulutusohjelma	Media Engineering
Ohjaajat	järjestelmäjohtaja Saara Dahlbacka tutkijaopettaja Aarne Klemetti

Insinöörityönä suunniteltiin ja toteutettiin digitaalisen journalistisen arkiston sisällön siirto ja yhdistäminen toisen arkiston kanssa. Julkaistujen dokumenttien arkisto takaa asiakasyrityksille tiedonhankintaan toimivan lähteen, jota voidaan helposti hyödyntää. Työhön kuului suunnitella osia käyttöliittymästä, valita ja yhdistellä metadatakentät, luoda metadatan tunnistava algoritmi sekä viedä arkiston sisältö toiseen arkistoon. Insinöörityössä perehdyttiin arkistoinnin periaatteisiin ja sovellettiin niitä käytännössä.

Insinöörityön tarkoitus oli parantaa siirrettävän arkiston käyttäjäystävällisyyttä ja modernisoida sen ulkoasu. Samalla arkiston metatietojen eheys tuli säilyttää. Arkistojen yhdistäminen ja lopputuotteen kehittäminen vaativat suunnittelutyötä sekä tietokantojen ja ohjelmoinnin ymmärtämistä.

Projekti järjestettiin yhdessä asiakasorganisaation työntekijöiden kanssa, jotka auttoivat muutamissa teknisissä asioissa. Projektin työskentely toteutettiin yleensä Skype-ohjelman kautta yhteistyössä järjestelmäasiantuntijan kanssa. Projektissa vaadittuihin ohjelmiin kuuluivat muun muassa vanhojen ja uusien tiedostorakenteiden hallintaohjelmat (Doris ja Trip), DbEdit ja UgEdit, joilla vanha arkisto voitiin sovittaa sulautumisen kannalta tarkoituksenmukaiseksi, ja DbTest tietokantojen hallintaan sekä Microsoft Office -ohjelmat.

Tuloksena syntynyt yhdistetty arkisto on toimiva ja säilyttää siirrettyjen dokumenttien metatiedot. Samalla se tarjoaa käyttäjille paremman käyttökokemuksen, selkeämmän esittelytavan arkiston sisällölle ja tarkemmat hakutoiminnot. Insinöörityössä suositeltuja periaatteita, kuten metadatakenttien kartoittamista ja valintaa sekä arkiston jaottelua käyttöliittymässä hyödyntämällä, voidaan saada aikaan toimiva arkisto vastaavien digitaalisten journalististen arkistojen yhdistämisessä.

Avainsanat	arkisto, digitaalinen arkisto, metadata, siirto

Contents

1 Introduction

These days media companies use a variety of programs for publishing. Different programs in turn create different workflows, practices, and ways to control documents in media companies. When media companies merge as the result of fusion or acquisition of share capital, work needs to be done to combine the different practices, workflows, and systems. Regarding these, usually the available options from both parties are examined and the best options are chosen and then applied in the whole company. Sometimes new systems, which the companies did not originally provide, are chosen.

Part of the workflow of media companies is archiving journalistic content. The proper archiving of published documents provides companies a reliable way to find information when creating new content. Access to the archived content can also be sold to customers which makes it a viable source of revenue for media companies.

Document control systems provide different ways to archive documents and not all systems are compatible from the beginning. Work must be done if incompatible systems are to be merged without loss of documents or metadata.

This thesis is a project documentation explaining the process of designing and carrying out the transfer and merger of the contents of a digital journalistic archive with another digital archive.

2 Archiving

2.1 Definition of Archiving

The word archive can mean several different things:

- a physical archive room
- an organizational unit or facility that upholds archives
- a set of documents created by the actions of the creator of the archive (Lybec 2006, 16).

Here, the creator of the archive means either an organization or a person, based on whose actions an archive is formed (Lybec 2006, 16).

In Finland, a law has been set about the forming and usage of archives. This mainly applies to government archives, but partly also to private archives. (Lybec 2006, 25.) Private organizations outside the jurisdiction of the law can define their own archive principles based on their needs. The archive law (831/1994, 6§) dictates that documents having either been delivered to or (been) formed by the actions of the creator of the archive, belong to the archive. The documents belonging to an archive can either be written or, for example, include graphic content or digitally produced content. (Archive law 831/1994, 6§.) It is up to the creator of the archive to design and take care of the archive (Archive law 831/1994, 8§).

A document in the context of an archive is one that contains information about its context (metadata), such as who has created it, when it has been created and for what purpose (Lybec 2006, 16). A document's most important attribute is its relation to the actions and tasks of the creator of the archive (Lybec 2006, 18).

The archive law (831/1994, 7§) dictates that archival work must maintain the usability of documents, take care of information services related to the documents, define the value of documents and remove unnecessary content from the archive. The requirements for archival work and preservation of documents must be taken into consideration when designing information systems. (Lybec 2006, 27.)

2.2 Digital Archiving and Metadata

Documents in digital form are easier to modify than conventional paper documents. Being in digital form, documents can be compiled with information taken from several systems. (Lybec 2006 13.) Yet, the forms and properties of digital documents create their own needs for handling them. For example, digital documents require proper programs for archiving them. (Lybec 2006, 70.)

An integral part of digital archiving is metadata. It is essentially information about information and an essential way of ensuring the context and validity of digital documents. Metadata makes searching from archives easier. In web publishing it makes information gathering easier. Lybec et al. (2006, 74) suggest that metadata becomes more significant when a digital document ages.

Metadata can exist in two ways: it can be internal and external (Lybec 2006, 73). The difference between these two should be recognized, as they direct the way an archive functions. As metadata can change while the actual digital file remains the same, the external way is to store metadata and files separately. For example, this can be done with an XML database and a separate file storage. This creates a flexible database but the link between the two files must be maintained by the creator of the archive. (Sharp 2007.)

The internal way is to create an object where the metadata and the file are stored together. As a benefit, this simplifies making backups and ensures that the files cannot become separated. However, the internal way can create quite large digital objects and, thus, makes their editing a complicated process. (Sharp 2007.)

When digital documents are transferred to new programs, new versions of software or new systems, special care must be taken to make sure the metadata attached to the files is kept intact. As Lybec points out, the format, or syntax, of metadata dictates whether it will transfer accordingly and in a usable form (Lybec 2006, 74). With broken, partial or non-existent metadata, an archive loses its usefulness.

Another aspect of digital archives are the user authorization rights. It is important not to allow unauthorized users access to edit the archived files. Only authorized users should

be able to edit metadata. These edits can include adding additional information or correcting spelling mistakes. Sharp (2007) notes that all metadata entries and editing should be recorded, so that it is possible to know which user did the changes, what changes were made and when did each change occur.

2.3 XML

XML (Extensible Markup Language) is a markup language similar to HTML (Hypertext Markup Language) in the way that tags define the structure of a document. A difference between them is that HTML is a fixed format, while XML is extensible. This makes XML a metalanguage. It can be used to describe other languages, letting a user create new markup languages. (Kyrnin 2017.)

The structure of a document is defined with the help of DTD (Document Type Declaration). DTD limits which elements can reside inside a certain element and what attributes they have. In this way, XML is especially useful for containing metadata in archives. Information in XML is structured and stored within nodes. Nodes can have one or several child nodes inside them. (W3Schools.)

Two kinds of XML databases exist: XML compliant and native XML databases. XML compliant databases can process XML documents but store information according to the document structure of the database. Almost all database manufacturers support XML data. Native XML databases use an XML document structure and withhold the physical structure of a document. (Tutorialspoint 2017.)

2.4 SQL

SQL (Structured Query Language) is a standardized query language developed by IBM. It allows users to make different kinds of searches, modifications and additions to a relational database. Virtually every relational database understands SQL. (Halvorsen 2016.)

SQL can be used to create, handle, and control databases. It defines structures and practices of cooperation for softwares. SQL servers are very versatile and can be used for a multitude of tasks, such as the database of a digital archive. (Chapple 2017.)

3 Merging of Two companies: Talentum and Alma Media

Talentum is a Nordic media company, mostly specializing in publishing magazines and books which are its most important products. It was founded in 1938 when the Talouselämä publication was founded. The company has since grown and during its lifetime it has acquired other publications under its name. The company also has other forms of revenue. These include publishing books, online services, professional training services, and events. The company has also spread from Finland to Sweden where there are similar operations taking place. (Talentum 2016).

The company's vision is to help professionals succeed in various work sectors, and thus, the target clients are from these sectors. All the areas of the company are working towards the same goal. The publications Talentum publishes are directed towards people in such areas of expertise as IT, stock market, and the housing sector, whereas other publications are directed for company leaders, medical staff, advertisers, and engineers. Published books also follow the same categories. Then there are the training services and events which are often held. They usually feature guest speakers who are well-known figures in their respected areas of expertise. An example of these is the Arvopaperi seminar for stock market enthusiasts. (Talentum 2016).

The structure of Talentum is divided into magazine publishing business (in Finland and Sweden), the event business, books publishing and legal training business, the direct marketing business and other functions. There are about 750 employees working at Talentum and they are spread across Finland, Sweden, Russia, Denmark, and the Baltic countries. (Talentum 2016).

Alma Media is a media and service company focusing on publishing and digital services. Their products include national, regional and local publishing operations, digital consumer and business services and printing and distribution business. The biggest and most well-known brands in the company include Kauppalehti, Iltalehti, Aamulehti, Etuovi.com, and Monster. Their international business operations are focused on recruitment services and marketplaces of business premises in Eastern Central Europe and Sweden. (Alma Media 2017).

Alma Media employs approximately 2,300 people (excluding delivery personnel). About one quarter of them work outside Finland. (Alma Media 2017).

On September 28, 2015 Alma Media acquired Talentum by purchasing all of its shares (Elo 2015). Moving forward with the merger, they formed a new business segment called Alma Talent (Alma Media 2016). The merging of the companies created the need to merge, among others, their IT systems. An integral part of the IT systems were the publishing platforms and digital archives.

At this point an opportunity rose to carry out a final year project by designing and implementing the transfer and merger of Talentum's archives with Alma Media's archive.

4 Programs Utilized in the Project

4.1 Doris32

Doris32 (henceforth referred to as Doris) is a 32-bit system for document control developed by a Finnish company, Anygraaf Oy. Doris utilizes an SQL database for its document management system. There are no set limitations to the stored documents; any kind of formats can be stored using Doris, such as text, image or spreadsheet formats. The documents stored in the database are opened and used with separate utility programs, chosen by the user. For example, text files can be opened with Microsoft Word, Notepad or any other text editor. The program also supplies its own text editor (Eddie) that can be used to produce content inside the software. (Doris32 user manual, 5.)

Each user has a specific view that is tied to their own username. The views are customizable by the program administrator. With selected tools in their own customized view, a user can control their content production. An example of a customized user interface with components, such as views and functions, as well as the database structures is shown in figure 1. All the documents and materials produced by the user are stored in the SQL database. Stored documents are recognized from the database using IDs. (Doris32 user manual, 5.)

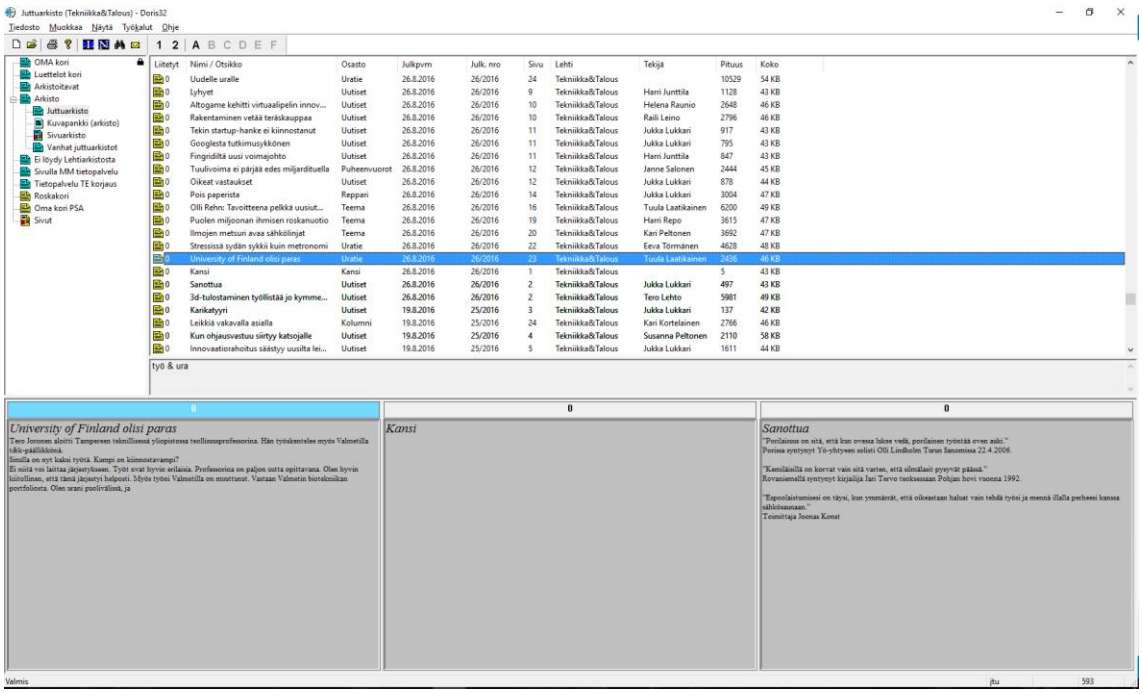

Figure 1. The text editing view of Doris, customized for archivers. Screenshot of Doris 2010.

Customized views as figure demonstrates 1 were utilized for publishing content. Both Talentum and Alma Media used Doris as the program for creating and storing the written content for their respective publications. Other formats, such as images, sheets, and graphs, were also stored in Doris before being sent to layout.

The program provides a comprehensive and customizable metadata input system allowing authorized users to add or modify metadata. Metadata is stored by Doris in separate XML files. Only if the/a document is produced with the program's own text editor, it is stored as one XML file containing both the written content and metadata. (Doris32 user manual, 72.)

Doris contains a search function to allow finding documents from the database. In addition to a normal database search, users can also search using timestamp, tags, classification, article type or free text search, provided the documents have proper metadata. More search criteria also exist and can be added based on the users' needs. (Doris32 user manual, 43.)

Talentum used Doris as the archiving system of the company for all published magazine content, as Doris provided the possibility for long term storing of content, metadata tagging, and search functions for accessing the archive.

The system also provides an inner command language and expendable SQL commands. In addition, Doris contains several separate control tools, such as a duty roster editor, a page planner program and a work planner program.

4.2 Tieto Trip

Tieto Trip (henceforth referred to as Trip) is an archive system originally developed in Sweden by Paralog AB. Since 1999 it has been owned and further developed by Tieto Finland Oy. Trip is a database system and a search engine. It is hybrid in that its search engine and database features cannot be separated. Depending on how it is used, Trip can be said to be a database system with an integrated search engine, or a search engine with database features (Bytespire Technology). This means that unlike Doris, Trip is solely used for archive purposes and cannot be used to produce content.

Trip uses a NoSQL database for its document storage. Content (such as documents, images, pages, and videos) is added to the database manually, meaning that it is inputted directly by the user inside the user interface or via constructed tunnels from linked software systems that input data directly via the blank Trip entry form, shown in figure 2. (Tieto 2017).

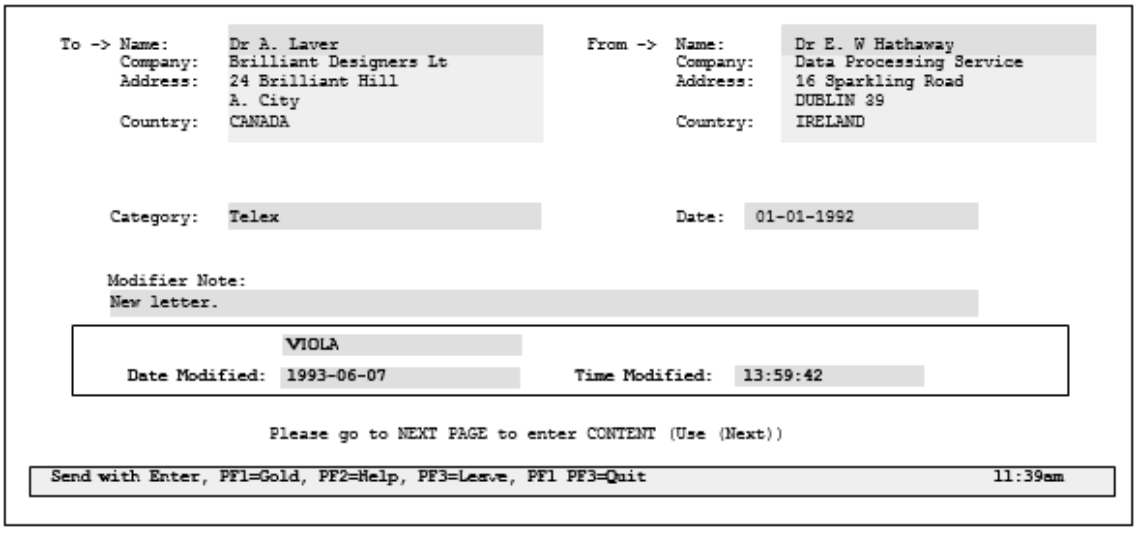

Figure 2. Data entry into database Corr, screen one of two. Copied from Tieto 2017.

Besides the above database entry form, the system allows for another way to input content into the system. Trip is an XML-enabled database, meaning that XML data is

mapped to and from an internal, non-XML storage structure upon storage and retrieval. XML data from existing documents can be transferred to Trip using TForm, which is a delimiter-controlled record format for the transfer of text into records intended for a TRIP database (Tieto 2017).

Trip provides a comprehensive and customizable metadata input system to add or modify metadata. Opening Trip, a user is given the option to access the system with or without logging in. Without logging in, users can only browse the archive. Modifying any documents or metadata requires users to login. (Tieto 2017).

Instead of using its own client program, Trip is accessed through a web browser. Using IP detection, it is inaccessible from outside a company's network. The archive opens to a blank view as seen in figure 3. In the view, users can choose what they want to search for.

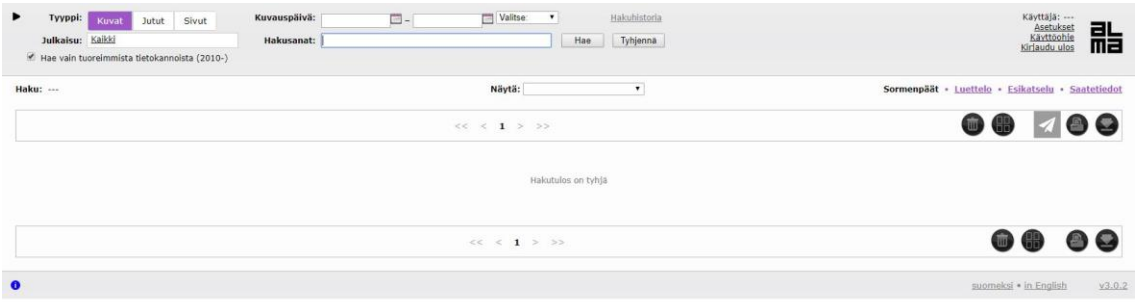

Figure 3. The opening view of the Trip archive system. Screenshot of Trip 2017.

From the opening view seen in figure 3, users can choose what type of documents they want to search for, and they can input the search criteria. Separate document type databases are selectable from the top of the page and the search can only be concentrated on one document type at a time. Trip has the capability to create a links between content types across databases. This means opening a document of one type shows links to

other content related to the original document. This makes finding all the relevant content regarding a published article easier.

Trip is used by Alma Media to store all their published content across all their magazine publications. The database and search functions of Trip can also be modified for other types of businesses besides media.

4.3 Newspilot

Newspilot is an editorial platform for planning, producing, and publishing content developed by a Swedish company, Infomaker Oy. Newspilot is a module-based program, meaning that the program contains separate parts for planning, writing text, multi-platform publishing, language checking, newsflow control, picture workflow, advertisement control and page publishing and printing. (Infomaker 2015.)

4.4 DbEdit and UgEdit

DbEdit and UgEdit are programs by Anygraaf Oy. They are used to edit the functions of Doris and customize it for different users. The names of the programs come from Database editor and Users & Groups Editor, respectively.

DbEdit that can be used to define settings for databases, tables and views used inside the Doris document control program. More specifically, it can be used to define which database and table a view opens and what search parameters are available to a user inside the view. DbEdit can also be used for editing scripts, defining conversions between databases and defining tags for the Doris editor and preview. (Db Edit Doris manual, 3). The opening view of DbEdit is shown in figure 4.

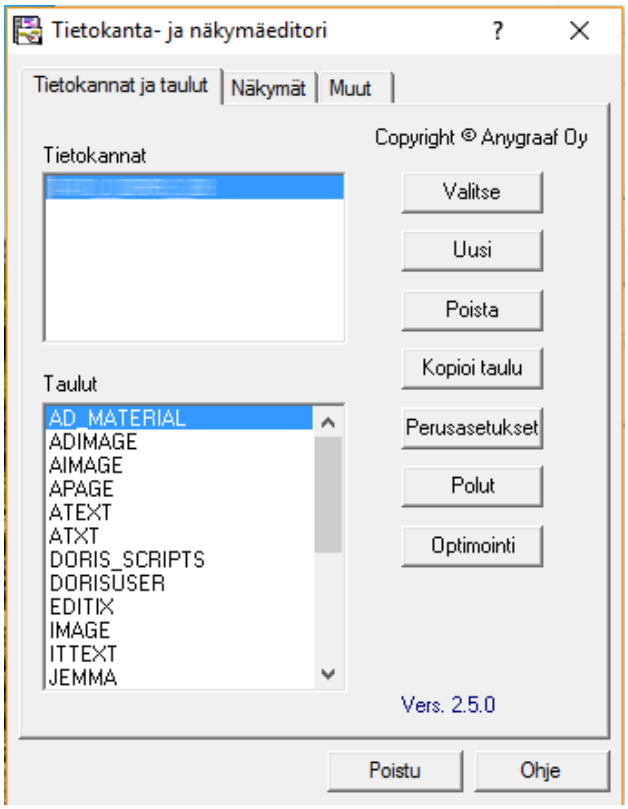

Figure 4. The opening view of DbEdit. Screenshot of DbEdit 2006.

Figure 4 shows the included buttons for creating, modifying or removing tables and views in the opening view of the editor. In the context of the project, DbEdit was used to modify certain tables' access search parameters to ease the archive export process.

The UgEdit program can be used to define user and group settings in Doris, such as which tables a user has access to, which groups a user belongs to and what their typical Doris view looks like (Ug Edit manual Doris, 3). In the context of the project, UgEdit was used to allow a certain Doris user account access to all the project's required tables.

4.5 DbTest

DbTest is a program used to execute SQL statements. The opening view of the program requires users to input the database location and login information. After login, the user is taken to the execute SQL window shown in figure 5.

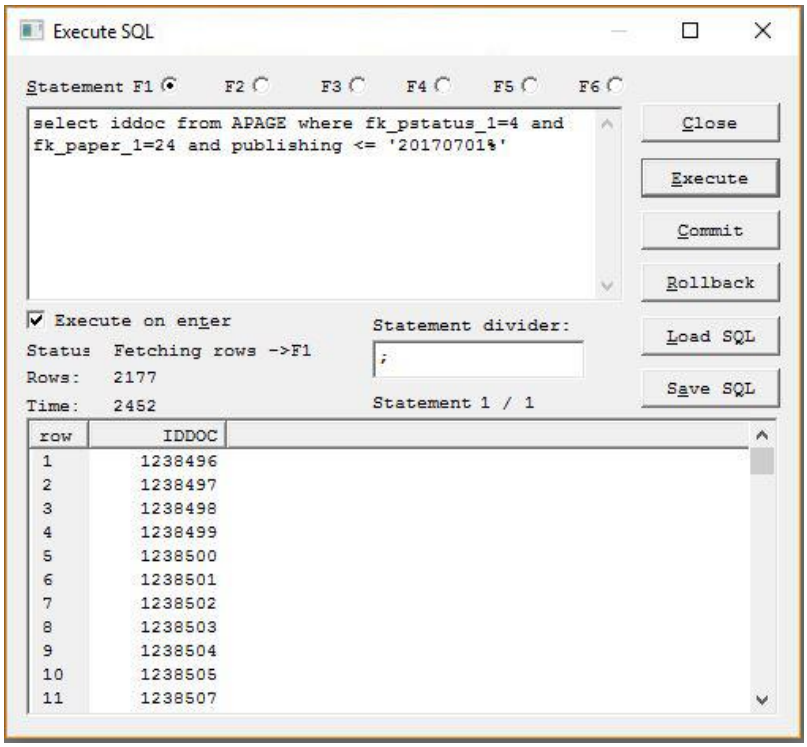

Figure 5. The execute SQL window of DbTest. Screenshot of DbTest 2006.

In the view of figure 5, users can input one or several SQL statements, load an existing statement or save them for later use, for example. The program was used in the project to extract ID lists of the exported documents from the database.

5 The Merger Project

5.1 Planning

The first phase of the project was to prepare a plan of how to approach the project. To not make the project too complicated, a simple solution was preferred with clear phases. The outline of the project was laid out into a flowchart presented in figure 6.

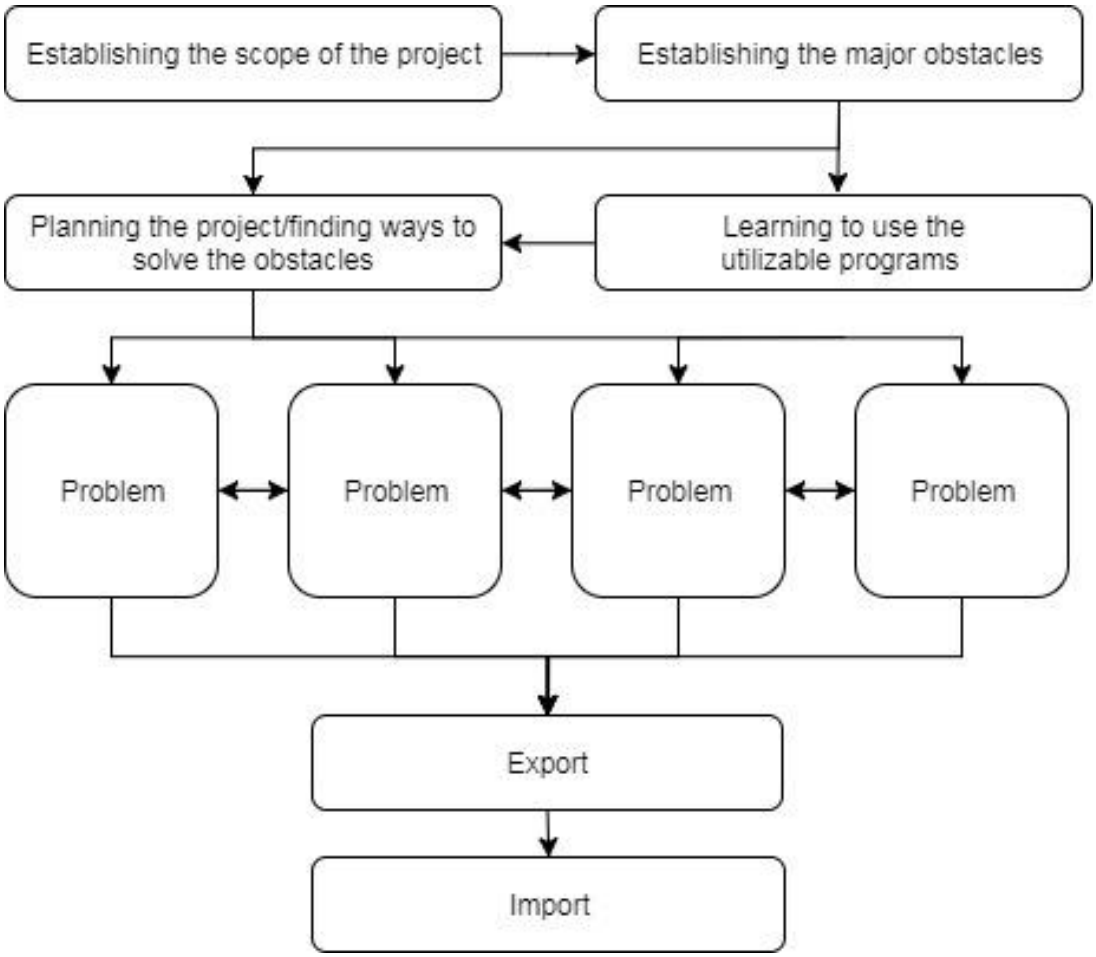

Figure 6. Flowchart of the project.

Figure 6 shows the initial idea of how to proceed in the project. After establishing the scope of the project and the problems to be solved, ways to solve these problems were sought and learning to use any programs that might be needed in solving them was focused on. The solving of the problems could be done simultaneously for a more efficient workflow. Finally, after all the problems had been solved, the project moved on to the final stages of exporting and importing of the archive.

The scope and requirements of the project to merge the archives of both Talentum and Alma Media were laid out in a meeting with my project instructor from the company, Saara Dahlbacka, a system manager. In the meeting, the following contents were outlined to be transferred from the Doris archive of Talentum (henceforth referred to as the Doris archive) and to be merged with the Trip archive of Alma Media (henceforth referred to as the Trip archive):

- The complete photo archive, containing an estimate of 270,000 photos in various file formats and sizes, complete with full metadata in the XML format. Estimated combined file space: 2 Tb
- The complete text archive, containing an estimate of 180,000 articles in the .lay format, complete with full metadata in the XML format. Estimated combined file space: 20 Gb
- The complete page archive, containing an estimate of 180,000 .pdf pages, complete with full metadata in XML format. Estimated combined file space: 200 Gb
- All non-archived photos, estimated to include 84,000 photos, with only partial metadata in the XML format. Estimated combined file space: 500 Gb
- The complete 'old text archive', early articles archived in several different file formats and with only partial or no metadata. Estimated combined file space: 100 Mb

Next, the problems and issues that would have to be solved during the course of the project were established:

- How the content's user rights will be handled in the new combined archive and how they will be transferred, keeping them visible to the user. How will the copyrights of photos be handled in the new combined archive?
- The Mediuutiset publication is partly owned by another media company and their image content is of sensitive value. The content was not to be visible to all users. How will the Mediuutiset publication's content be accessible in the new combined archive?
- How is the metadata regarding previous usage of photos transferred correctly to the new system?
- As Alma's PDF files are saved in a different way, do archived PDF files from the Doris archive need processing so that they will show up correctly in the Trip archive. The Trip archive has a built-in PDF viewer for single pages. Will multi-page PDF files function properly in the new archive?
- From a UI perspective, how will the Doris archive's contents be presented next to Alma Media's content in the new archive system?

After the aforementioned problems would be solved, the earlier outlined contents of the Doris archive were to be exported into a removable hard drive. There, the contents would be organized into an appropriately named folder structure. The hard drive would then be

delivered to Timo Kiviniemi, a system specialist, who would then import the contents into the Trip archive. (Saara Dahlbacka 2016.)

A few more meetings were held with people from Alma Media joining through Skype. The people who joined the meetings were Ville Inkinen, a system specialist, Timo Kiviniemi, and Kari Hurtola, a development manager. Since Alma Media has used the same Doris document control system before, they had more knowledge of how to proceed with the archive merger in technical terms. In these meetings the following points regarding priorities in the merger were established: (Dahlbacka; Inkinen; Kiviniemi & Hurtola 2016.)

- The old text archive and all non-archived photos would be transferred last, due to their low usage and priority.
- Talentum has saved all PDF pages in two qualities: print quality and web scaled quality. Both sets of PDF files would be stored into the Trip archive. Web scaled pages are contained in the Doris archive and would be transferred normally, whereas the print quality PDF pages are stored on separate disks. They will be imported at a later date as a separate project.

5.2 Rights Issue

Meetings were held with Timo Pylvänäinen, head of photography, where the question of how the user rights of the photos from the Doris archive should be presented in Trip was discussed. This was necessary to avoid any misuse of potentially sensitive images. The transfer of the photo archive required the examination of these rights. At the time Talentum was active, the archived photos had different permissions of use and were divided into four categories, which are listed below:

- Corporate rights. Approximately 70 % of the photos are tagged with these rights.
- Publication rights. Approximately 20 % of the photos are tagged with these rights.
- One-time rights. Approximately 9 % of the photos are tagged with these rights.
- Other rights. Approximately 1 % of the photos are tagged with these rights.

Alma Talent's Trip archive calls rights as "limitations" and natively provides two different options: "only for own use" (equal to publication rights) and "one-time rights". The Trip archive is flexible and also allows for other types of rights to be inputted into the metadata

field. Using this feature, the existing rights from the Doris image archive could be transferred into the new system, if the metadata field containing them could be straightly transferred to the corresponding one in the Trip archive metadata field.

The larger user base of the Trip archive presented problems with too many users having access to the contents. A way was conceived with Mr. Pylvänäinen - a type of content based lock mechanism to prevent the accidental re-use of selected photos (Pylvänäinen 2016). This was later confirmed to be a possible add-on to the Trip archive by Ville Inkinen, a system specialist.

As further inspection of the copyright issue is outside the scope of this thesis, it will not be further expanded upon. What can be said is that the copyrights of the photo archive's contents could be transferred to the new system without compromise (Talentum avustajasopimus) (Pylvänäinen 2016).

5.3 User Experience of Trip

The next step was to design the way the Doris archive's contents would be presented to the user in Trip. Naturally, the same UI of Trip would be used for browsing the archive. The initial vision of how to locate and focus searching on the imported contents of Doris was to add every Talentum publication to the list of Alma publications in Trip, as seen in figure 7. The publication values to generate the list would be extracted from the metadata of the contents.

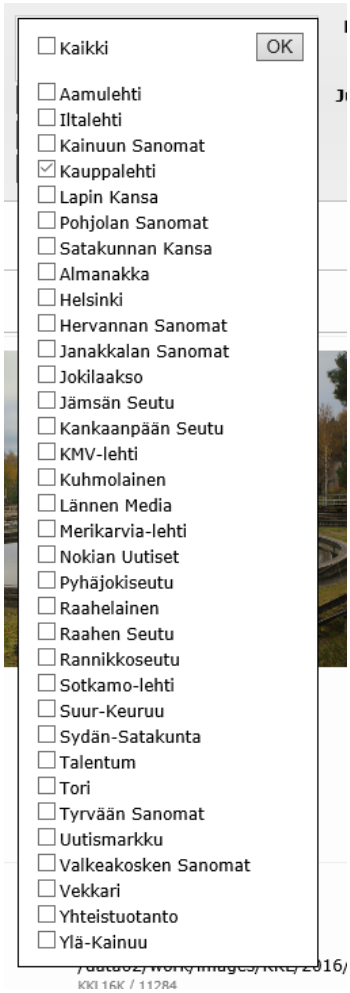

Figure 7. List of publications in Trip. Screenshot of Trip 2017.

As the list of Talentum publications to be added was quite long, it would have made the list even longer and unuseful from a UX point of view. Another point which would extend the list was that new content under the same publication name would be shown as a separate entry. This was due to the content arriving from Newspilot and being saved in a different root in the database.

Working together with Timo Kiviniemi, he suggested saving all the contents of the Doris archive inside the same root in the Trip database. From a user standpoint, this would create one new selectable entry into the publication list, and after the initial search, the results could be narrowed down using a drop-down menu. The menu would contain the list of Talentum publications imported from the Doris archive and the same publications published under Alma Talent. This drop-down menu can be seen in figure 8.

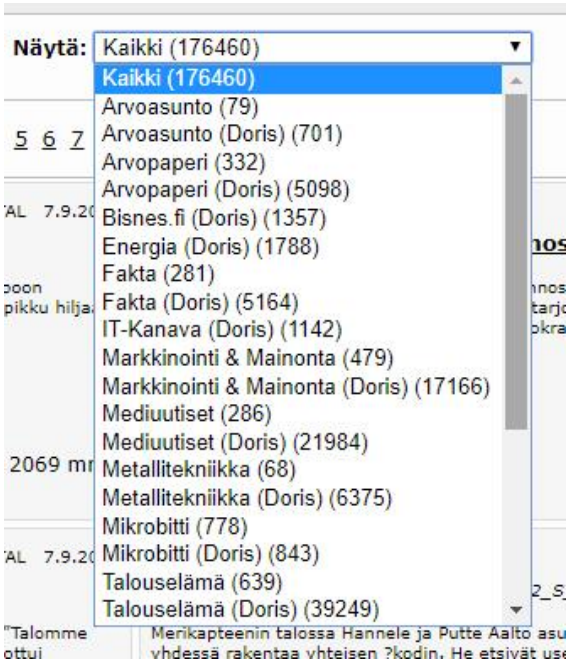

Figure 8. The drop-down menu. Screenshot of Trip 2017.

As can be seen in figure 8, the content imported from the Doris archive can be identified with the notation ‘(Doris)’ and the ones without it are new content imported through the Newspilot program. The root entry in the publication list was named ‘Talent’. The idea is to add possible new publications to the Alma Talent roster into the same ‘Talent’ root in the database so they will be visible to users in the same list.

As the Mediuutiset publication’s photos were supposed to be visible only to the personnel of the publication’s editorial team, the decision was made to export them into a different root in the database. This root would only be visible if the Trip archive was accessed with certain login credentials handed out to selected editorial team members.

All non-archived photos and the old text archive would receive a similar process. Login credentials would be handed out to photo journalists and information services personnel, respectively.

Multi-page PDF files from the Doris archive were tested to make sure they work in Trip. They did not show up properly using the Trip PDF viewer but were still functional enough to be accessible and usable. As the amount of PDF files to break down into single-page files was too big, the decision was made to leave them as they were. Timo Kiviniemi informed me that modifications will later be done to Trip’s PDF viewer to make sure the

multi-page PDF files show up properly. Once the modifications had been done, the PDF files showed correctly in the viewer, as seen in figure 9.

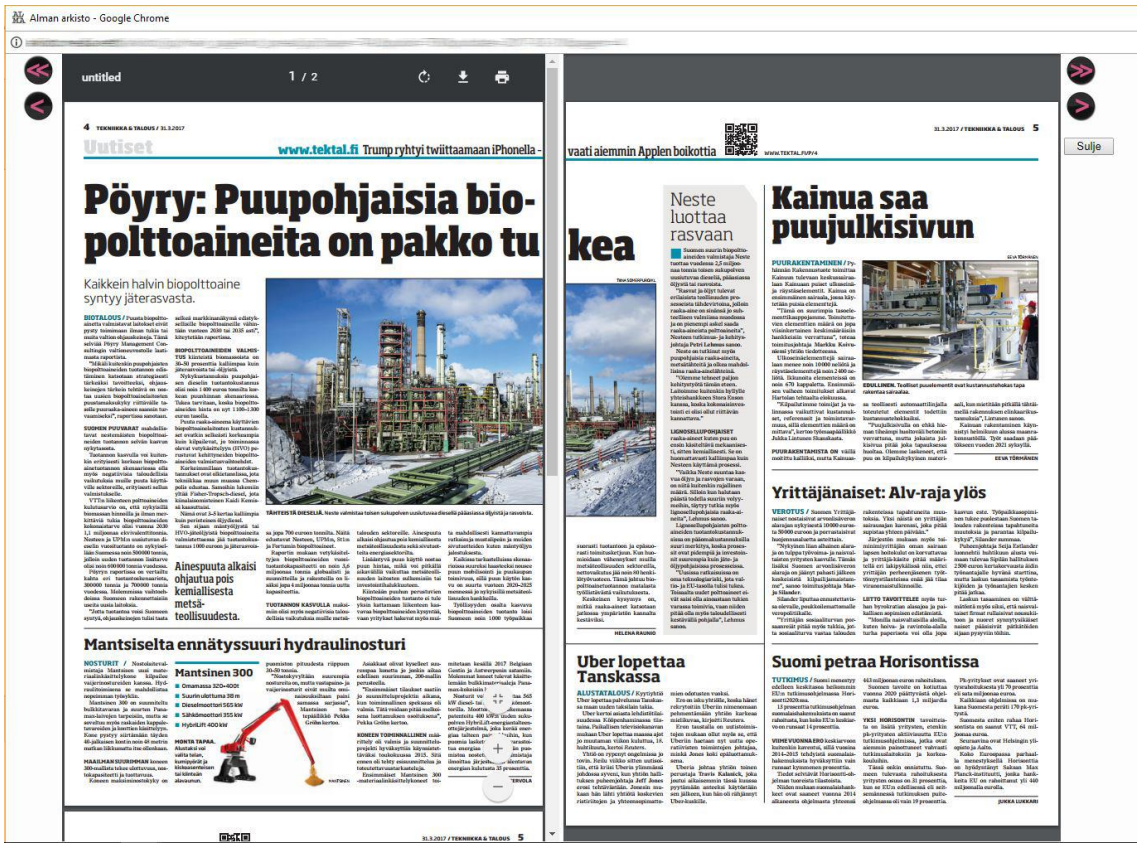

Figure 9. The Trip PDF viewer. Screenshot of Trip 2017.

Trip's functionality of linking images and pages to a written article was to be taken into use. The imported Doris documents contained the required metadata to link them. At the time the project was ongoing, this feature was not functioning properly. Timo Kiviniemi assured that the functionality will be fixed at a later time and that the links to related documents will appear.

5.4 Metadata

As mentioned earlier, when digital documents are transferred to new systems, special care must be taken to make sure the metadata attached to the files is kept intact. With this in mind, the metadata fields from the different document types in Doris archive were listed in order to find and allocate the corresponding metadata fields in Trip. All metadata would be transferred from the documents in Doris to Trip. As Trip either missed, provided

more or had differently named metadata fields for documents than Doris, some reallocation had to be done and possible wastage could occur. The metadata fields' importance had to be evaluated at this stage.

Evaluation of the metadata was based on three factors:

- Was the content created by users or automatically created by the system?
- Is the metadata field's information content useful in Trip?
- Does the metadata field have a matching field in Trip?

Based on these factors important metadata fields were picked. These were then allocated to the corresponding Trip metadata fields in Skype meetings with Timo Kiviniemi. Table 1 illustrates the Doris to Trip metadata field allocations in all three document types.

Table 1. The Doris to Trip metadata conversion.

Images			
Doris metadata field	Trip metadata field	Doris metadata field	Trip metadata field
Käsittelyohjeet	Huomautus	Julkpvm	Julk.historia – Julkaisupäivä
Nimi	Kuvan nimi	Kuvausaika	Kuvauspäivä
Kuvaaja	Kuvaaja	Internet	**Not used**
Kuvauspaikka	Paikka	prev_id	**Not used**
Osasto	Luokka 1	prev_name	**Not used**
Lehti	Julk.historia – Julkaisu	Luotu	Luontipäivä
IPTC	**IPTC (Not used)**	Käsitelty	**Not applicable**
Kori	**Not used**	Saatetiedot muutettu	Muutospäivä
Tyyppi	**Not used**	Saatetietojen muuttaja	**Not applicable**
Käsittely	**Not used**	Id	Kuvan ID
Oikeudet	Rajoitukset	Käsittelijä	**Not applicable**
Status	**Not applicable**	Tyyppi	**Not applicable**
Julk. nro	**Not applicable**	Luoja	**Not applicable**
Käsittelykoko	**Not used**	Kuvaus	Sisältö
		Asiasanat	Avainsanat
Text			
Doris metadata field	Trip metadata field	Doris metadata field	Trip metadata field
Nimi / Otsikko	Otsikko	Internet	**Not used**
Osasto	Osasto	Rooli (WWW)	**Not used**
Käsittelijä	**Not applicable**	Luotu	Luontipäivä & Luontiaika

Tekijä	Kirjoittaja	Käsitelty	**Not applicable**
Lehti	Julkaisu	Saatetiedot muutettu	**Not applicable**
Kori	**Not used**	Saatetietojen muuttaja	Muuttaja
IPTC	IPTC	Id	Jutun ID
Status	**Not applicable**	Käsittelijä	**Not applicable**
Sivunimi	**Not applicable**	Tyyppi	**Not applicable**
Julk.aika (WWW)	**Not used**	Luoja	**Not applicable**
Oikeudet	**Not applicable.**	Kuvaus	Esirivi
Julk. nro	**Not applicable**	Asiasanat	Asiasanat
Julkpvm	Julkaisupäivä	Linkit	**Not used**
Pituus	Pituus	Luokittelu	**Not applicable**
Sivu	Sivu		
PDF			
Doris metadata field	Trip metadata field	Doris metadata field	Trip metadata field
Pohja	**Not used**	Ad Ratio	**Not used**
Nimi	**Not applicable**	Sivuja	**Not used**
Sivunimi	**Not used**	Luotu	**Not applicable**
Osasto	**Not used**	Käsitelty	**Not applicable**
Teema	**Not used**	Saatetiedot muutettu	**Not applicable**
Lehti	Julkaisu	Saatetietojen muuttaja	**Not applicable**
Tekijä	**Not used**	Id	Sivun ID
Käsittelijä	**Not used**	Luoja	**Not applicable**
Status	**Not applicable**	Käsittelijä	**Not applicable**
Deadline	**Not used**	Tyyppi	**Not applicable**
Sivunumero	Sivu	Kuvaus	**Not used**
Julkpvm	Julkaisupäivä		

In table 1, the cells marked with ‘Not used’ mean that the information in these metadata fields was not utilized. The fields were either blank and or contained information automatically created by the system. Examples of these include the type and internet usage metadata fields that were not used by Talentum.

The cells in table 1 marked with ‘Not applicable’ mean that the metadata contained in the field was not critical information and that a corresponding metadata field for it was not present in Trip. An example of this is the publication number. It was used in Doris to denote the issue of a publication in which the document was published. Trip instead uses the publication dates to present the same information and has no field for publication numbers. Thus, the publication number was not carried over from Doris to Trip.

During Skype meetings, Timo Kiviniemi pointed out that there are hidden metadata fields in both Doris and Trip that are not visible to users but which can be used as search parameters. One such metadata field was the IPTC classification. Doris documents contained this field, even though it was not used to my knowledge. A decision was made to carry it over to Trip as to not lose any possible metadata that might exist within the documents.

After the metadata fields from Doris were confirmed to allocate to the correct counterparts in Trip based on the design of table 1, the conversion was programmed into the Trip import process by Timo Kiviniemi. After that, test files were sent to him so that he could check the correctness of the import process. Small errors were found that caused the import process to fail. These were being caused by for example empty date fields in some of the Doris metadata files. These empty date fields did not cause errors in Doris but they would in Trip. A decision was made to replace the empty fields with an 'empty' date. This change was then made into the import process coding to bypass the errors and to automatically fill the empty metadata fields.

The 'Rajoitukset' ("Limitations") metadata field is to be replaced at a later date with a placeholder text to denote the publisher inside Alma Media. This can be simply done with an override command for all the files after they have been imported.

5.5 Information on Previous Usage

During the time the Doris archive of Talentum was in use, the re-use of images from the archive was recorded into metadata. The systematic marking of checking and adding this metadata was part of the archiving process. As the Doris metadata fields did not include a specific field for this purpose, the information was written in the free text field, among the associated tags and keywords of the image. Whenever an image that was extracted from the archive appeared in a published magazine, the previous appearance of the image was documented in the metadata free text field of the image, and the publication value and publication number metadata fields were changed to reflect the new published appearance.

A specific syntax was used for this marking which allowed users to find the re-used images from the archive using the free text search. About 6,000 images contained this syntax in their metadata.

The syntax that was used was in the following format:

- KÄYTETTY AIKAISEMMIN (“USED BEFORE”), written in all caps
- the shortened name of the publication, written in all caps
- the number of the publication
- if more publications exist, they are written next, separated by a comma

Here is an example of the syntax:

```
“KÄYTETTY AIKAISEMMIN TE 12/2010, 24/2011, AP 5/2012”
```

The way the syntax is placed among tags and keywords in the free text field can be seen in figure 10.

Figure 10. The information of previous usage in the Doris free text metadata field. Screenshot of Doris 2010.

In the Trip archive system there is a specific metadata field for the image’s previous usage. This field can be seen in figure 11. The notable difference is that the Trip metadata field uses publishing dates instead of publishing numbers and whole names for the publications instead of shortened ones.

Figure 11. The information of previous usage in the Trip metadata field. Screenshot of Trip 2017.

The information of previous usage needed to be transferred to the Trip field seen in figure 11 from the free text field of figure 10. Discussing the problem with Timo Kiviniemi, a

possible way of transferring this metadata correctly was discovered. If the metadata of previous usage were to be situated in the XML inside a separate node, it could be picked up by the import process of the Trip archive. Timo Kiviniemi explained that the modification to the import process could be made easily.

A structure of the new XML node ‘publishinghistory’ was designed and can be seen in listing 1.

```
<publishinghistory>
     <publishing>
          <pubdate>DATE OF PUBLICATION</pubdate>
          <publication>PUBLICATION NAME</publication>
     </publishing>
</publishinghistory>
```

Listing 1. Structure of the ‘publishinghistory’ XML node

This XML node structure would reside inside the root node ‘dorisdocument’ and be placed after the ‘historyinfo’ child node. The new XML node was then presented to Timo Kiviniemi, who approved it and made the modification to the Trip import algorithm to account for it.

The Doris database included tables that contained the corresponding dates of the publishing numbers and the whole names of the shortened publication names, as seen in appendix 1. Using that information, the metadata from the syntax would have to be substituted and then placed inside the ‘publishinghistory’ node according to the structure.

An algorithm was designed to carry out the aforementioned actions. The algorithm works as follows:

1. Inside the XML of the image, identify the string of characters beginning with “KÄYTETTY AIKAISEMMIN” in the description node.
2. Create a child node called ‘publishinghistory’ inside the root node in the XML. ‘publishinghistory’ will contain one ‘publishing’ child node which contains the child nodes ‘pubdate’ and ‘publication’.
3. Identify the next part of the string by comparing it to the list of short names of publications.
4. Compare the recognized publication short name to the list of full names of publications and insert the correct publication full name into the ‘publication’ node.
5. Identify the next part of the string: Check the publication number, format “nn/yyyy”.

6. Compare the format to the list of publication numbers and their corresponding dates.
7. Insert the correct date in the “yyyymmdd” format into the ‘pubdate’ node.
8. Continue checking the string.
9. If the next part of the string is a publication number:
    a. Create a new ‘publishing’ node
    b. Insert the same publication full name to the ‘publication’ node as in the previous node
    c. Move to step 6 and continue
10. If the next part of the string is a publication number:
    a. Create a new ‘publishing’ node
    b. Move to step 4 and continue
11. Otherwise, stop checking the string.

A request for quotation was made to Anygraaf Oy to use the above algorithm and to convert the roughly 6,000 images’ XML files to include the modified information on previous usage. This work request can be seen in appendix 2. The sub-project of converting the XMLs was carried over remote desktop control to the Doris servers with me opening the access to an Anygraaf employee.

Along the way, some problems appeared. Some of the ‘used before information’ -syntaxes in the metadata were not correctly inputted and caused errors with the algorithm. Luckily, the amount of incorrect syntaxes was relatively small and they could be corrected by hand. In the end, the sub-project was successful and took about a month to complete. The new XML test files were sent to Timo Kiviniemi who used them to check that the import process worked correctly.

5.6 Export

As the initially stated problems were solved, the transfer was ready to begin. Saara Dahlbacka orientated me into using the DbEdit and UgEdit programs which I then used to create the Doris view suitable for exporting. This view can be seen in figure 12.

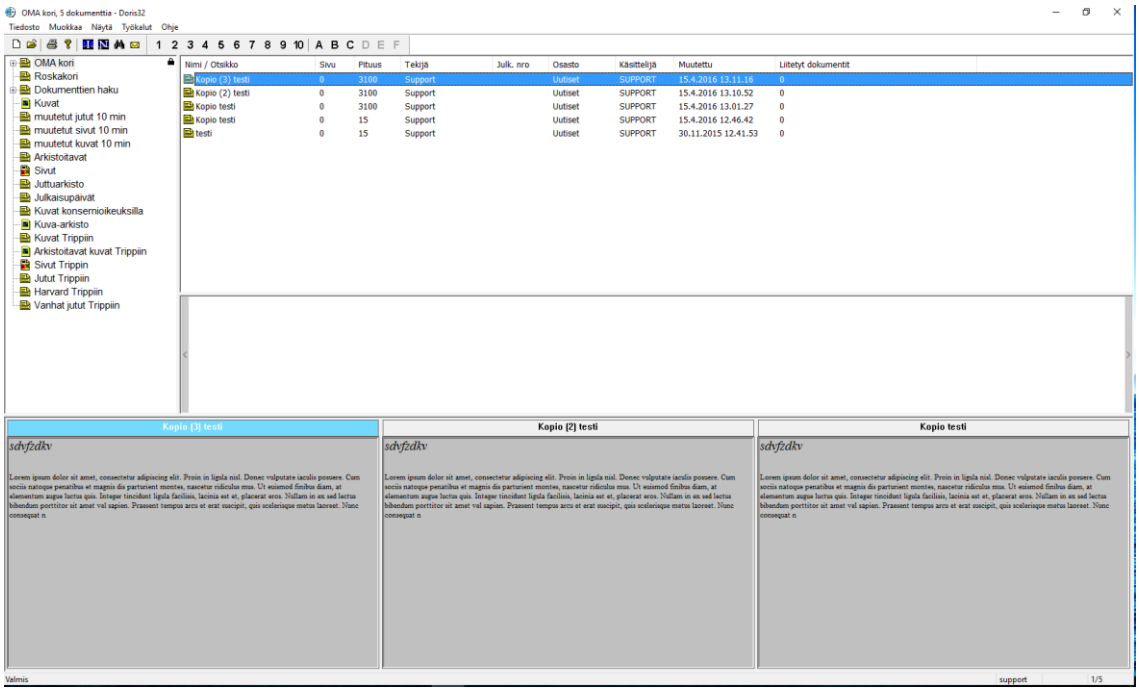

Figure 12. The export view of Doris. Screenshot of Doris 2010.

The required views were for the picture, photo and page archives. The un-archived photos and the old text archive are visible on the left in figure 12. The search parameters for these were set to contain free text search, selectable publication value and the range of creation date.

The NAS external drive was coordinated to be delivered to the server room and then connected into the network. After it became visible over the network, the paths of the drive were programmed into the export functions of Doris. These functions can be seen as the buttons labeled 1-10 in the top bar of the view in figure 11. Tests were run to measure the approximate amount of time needed for the export. The tested export times and total export time approximations are presented in table 2.

Table 2. Export timetable

Tested export times per 1000 files	
Image files	5 minutes 19 seconds
PDF files	3 minutes 52 seconds
Text files	1 minute 30 seconds
Estimated export times	
Archived image files	24 hours
Unarchived image files	7 hours
PDF files	12 hours
Text files	4,5 hours

The export had to be done in sections of a few thousand files at a time. This was done to avoid overloading or crashing the servers, as at the time they were still also used for editorial work. Exporting a small number of documents at a time was also a good way to prevent export errors. As the exporting was done during office hours and amongst other jobs, the estimated export times were prolonged. As previously mentioned, the un-archived images were left out of the transfer at this point. The total time for the export was about two work weeks. After the export was complete, the NAS external drive was delivered to Timo Kiviniemi and then connected into the systems in the receiving end.

5.7 Import

Timo Kiviniemi started running the import process for the Doris documents off the NAS external drive, but a problem appeared that prevented the import.

While exporting the contents of the archives to the removable hard disk drive, I made an assumption that the computer and system in the receiving end would be able to read the contents the same way as the computer I used to export the contents. Thus, all the contents of a single document type (image, text, and page) were exported to a single folder and no subfolders for different publications were created. Once the external drive was connected to the receiving computer, the Linux based queries to the folder's contents failed as the amount of data was too massive to parse.

In a meeting with my instructor Saara Dahlbacka, she suggested to use SQL queries to pull lists of IDs from the Doris database. These lists could be limited to contain a specific

publication's contents and from what time period they were from. Saara Dahlbacka provided instructions on how to use the DbTest program to extract these ID lists. She also provided the basic SQL query from which further queries for different archive types were derived.

I then combined the extracted IDs with a basic structure of a UNIX command prompt provided by Timo Kiviniemi in Microsoft Excel, as seen in figure 13.

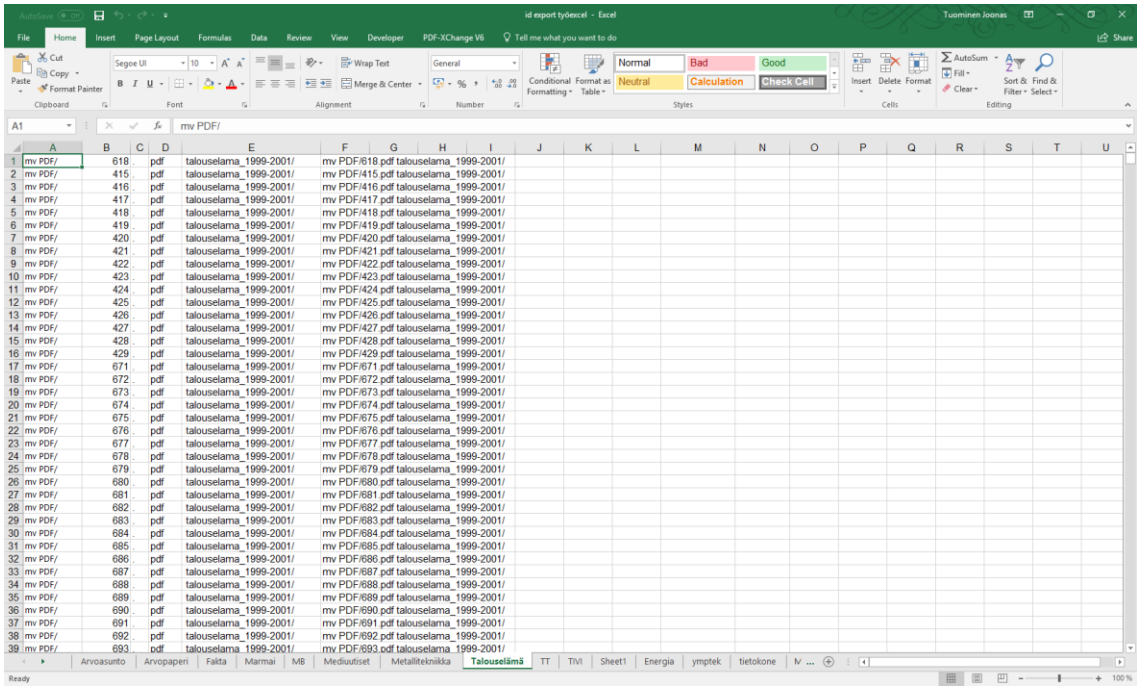

Figure 13. ID lists being duplicated in Excel. Screenshot of Microsoft Excel 2016.

Using Excel's fill handle functionality, the command prompt could be duplicated for each individual file instantly. These lists of command prompts were then sent to Timo Kiviniemi in TXT files via email. He could then run the commands instantly, dividing the mass of files into smaller, dedicated folders for each publication. The biggest publications were also divided into subsets containing a range of a few years.

After the documents were divided into subfolders, the import process started working and the documents started to appear in Trip. Some errors appeared during the import process that were due to some document files being 0 kilobytes in size. A check was made for those documents to locate them in the Doris archive. Each document was located and confirmed to be 0 kilobytes in the Doris archive, ruling out the possibility of an export/import error. The files were deemed corrupt and were left out of the transfer.

The whole import process took about three months, as again it was done during office hours and amongst other jobs and was carried out over the summer when key personnel regarding the process were on vacation.

6 Conclusion

As pointed out in the introduction, this thesis focused on the project of transferring and merging the digital archives from one archive system to another. Research was first done into good archiving practices and the ways to maintain the integrity of the archive. The project was then planned with people joining from the client company to aid in the technical details. Working mostly over Skype with a system specialist, a user interface was created and metadata transfer and exporting of the documents were conducted successfully and as planned.

The client company benefitted from the project. The user experience of using the archive is now much better as the user interface of the new archive provides significant improvements over the user interface of the previous archive. Users no longer need a separate program to access the archive and the new login system prevents accidental editing of metadata. Browsing between documents is considerably easier as the search results are displayed in a more structured way and the imported documents appear equal next to other documents in the archive.

The number of documents that were successfully transferred to the Trip archive was almost 100% of the total documents in the Doris archive. Only corrupt documents were purposefully not transferred.

Portions of the project took too much time due to human errors. If done again, a tighter interaction with other project members would minimize any misunderstandings. Still, the project was well received in the company. The development of the archive will continue within the company with emphasis on user experience and solving the problems mentioned in the thesis. The documentation provided in this thesis can be benefitted from if there is a need to merge digital archives.

References

Alma Media. About us [online]. Finland, Alma Media.
URL: http://www.almamedia.fi/en/about-us/this-is-alma.
Accessed 1 February 2017.

Alma Media. Alma Median ja Talentumin yhdistymisessä syntyvän uuden liiketoimintayksikön nimeksi Alma Talent [online]. Finland, Alma Media; 5 February 2016.
URL: https://www.almamedia.fi/uutishuone/uutinen/05-02-2016-alma-median-ja-talentumin-yhdistymisess%C3%A4-syntyv%C3%A4n-uuden-liiketoimintayksik%C3%B6n-nimeksi-alma-talent.
Accessed 12 November 2017.

Bytespire Technology. Products: Tieto TRIP [online]. Finland, Bytespire Technology.
URL: http://www.bytespire.com/products_tietotrip.php.
Accessed 1 March 2017.

Chapple Mike. SQL Fundamentals [online]. ThoughtCo; 22 July 2017.
URL: https://www.thoughtco.com/sql-fundamentals-1019780.
Accessed 5 August 2017.

Dahlbacka Saara. System Manager, Alma Talent, Helsinki. Personal communication 3 October 2016.

Dahlbacka Saara, Inkinen Ville, Kiviniemi Timo & Hurtola Kari (2016). System Manager; System Specialist; System Specialist; Development Manager, Alma Talent, Helsinki. Personal communication 14 October 2016.

DbEdit [computer program]. Version 2.5.0. Finland, Anygraaf Oy; 2006.
Accessed 21 March 2017.

DbTest [computer program]. Version 2.5. Finland, Anygraaf Oy; 2006.
Accessed 21 March 2017.

Db Edit Doris manual [e-document]. Finland, Anygraaf Oy.
Accessed 22 March 2017.

Doris [computer program]. Version 2.6.4.212. Finland, Anygraaf Oy; 2010.
Accessed 21 March 2017.

Doris32 user manual [e-document]. Finland, Anygraaf Oy.
Accessed 12 February 2017.

Elo Emil. Alma Media ja Talentum yhdistyvät [online]. Helsinki, Finland, Kauppalehti; 29 September 2015.
URL: http://www.kauppalehti.fi/uutiset/alma-media-ja-talentum-yhdistyvat/6tc2FxyZ.
Accessed 1 February 2017.

Finlex. Archive law 831/1994 [online]. Finland, Finlex.
URL: http://www.finlex.fi/fi/laki/ajantasa/1994/19940831.
Accessed 2 March 2017.

Halvorsen Hans-Petter. Structured Query Language [online]. Norway, University College of Southeast Norway; 8 January 2016.
URL: http://home.hit.no/~hansha/documents/database/documents/Structured%20Query%20Language.pdf.
Accessed 10 October 2017.

Infomaker. Newspilot 4 Reference Guide [e-document]. Sweden; 2015.
URL: https://wiki.infomaker.se/display/NPM42ENG/Newspilot+4+Reference+Guide.
Accessed 11 February 2017.

Kyrnin Jennifer. What are Markup Languages [online]. ThoughtCo; 5 July 2017.
URL: https://www.thoughtco.com/what-are-markup-languages-3468655.
Updated 5 July 2017.
Accessed 10 October 2017.

Leinonen Heikki. Alman arkisto – Käyttöohje [e-document]. Helsinki, Tieto Finland Oy; 2011.
Updated 29 October 2014.
Accessed 1 March 2017.

Lybec Jari, Pirilä Pirkko, Rosberg Harri, Vappula Jorma, Rosberg Harri, Selin Rauno & Leppänen Markku. Arkistot yhteiskunnan toimiva muisti. Asiakirjahallinnon ja arkistotoimen oppikirja. Arkistolaitos, Helsinki; 2006.

Microsoft Excel [computer program]. Version 1708. Microsoft; 2016.
Accessed 3 November 2017

Pylvänäinen Timo. Head of Photography, Alma Talent, Helsinki. Personal communication 21 October 2016.

Sharp Robert. Solving archive challenges [online]. Research information; 11 June 2007.
URL: https://www.researchinformation.info/feature/solving-archive-challenges.
Accessed 3 September 2017.

Talentum [online]. Finland, Talentum.
URL: http://www.talentum.com/fi/company/.
Accessed 27 January 2016.

Talentum Freelancer contract. Helsinki, Talentum Oy.

Tieto. Tieto TRIP Wiki [online]. Finland, Tieto; 5 May 2017.
URL: https://trip.service.tieto.com/wiki/dashboard.action.
Accessed 1 July 2017.

Trip [computer program]. Version 7.1-3.1. Finland, Tieto Oy; 2017.
Accessed 16 October 2017.

Tutorialspoint. XML – Databases [online]. Tutorialspoint; 2017.
URL: https://www.tutorialspoint.com/xml/xml_databases.htm.
Accessed 10 October 2017.

Ug Edit manual Doris [e-document]. Finland, Anygraaf Oy.
Accessed 22 March 2017.

W3Schools. XML DOM Nodes [online]. W3Schools.
URL: https://www.w3schools.com/xml/dom_nodes.asp.
Accessed 10 October 2017.

Publication list of short names and full names

- AP (Arvopaperi)
- AA (Arvoasunto)
- ENERGIA/EN (Energia)
- FA/F (Fakta)
- MB (Mikrobitti)
- MM/M&M (Markkinointi&Mainonta)
- MPC (MikroPC)
- MT (Metallitekniikka)
- MU (Mediuutiset)
- TE (Talouselämä)
- TT/T&T (Tekniikka&Talous)
- TV/TIVI (TIVI)
- TK (Tietokone)
- TH (Tekniikan Historia)

The request for quotation to Anygraaf Oy

Tarjouspyyntö ja työarvio 10.1.2017, päiv. 20.1.207

Työ:

Kyseessä on Doris-kannan Image-taulu ja XML:t ovat kuvien saatekaavion tietoja.

Tarkoitus on siirtää XML:ssä olevassa elementissä oleva tekstimuotoinen sisältö toisen noden sisältämiin elementteihin sisällöksi sekä konvertoida tekstissä olevat lehtiarvot lyhenteistä kokonaisiksi ja julkaisunumerot päivämääriksi.

Kuvia on n. 6000 kpl.

Lähtötilanne:

Doris-arkiston kuvien XML:ssä on description-elementin sisällä tekstiä. Tämän tekstin joukossa esiintyy tietyllä formaatilla olevan stringi.

Stringi on aina muotoa

KÄYTETTY AIKAISEMMIN (lehtiarvo) (ilmestymisnumero)

Lopputilanne:

Stringistä saatu tieto on jäsennelty uuteen nodeen ja sen sisälle elementteihin

```
<publishinghistory>
      <publishing>
        <pubdate>ILMESTYMISPÄIVÄMÄÄRÄ</pubdate>
        <publication>LEHTIARVO</publication>
      </publishing>
</publishinghistory>
```

Huomioita:

1. Stringissä olevat lehtiarvojen nimet tulee korvata nodeen kokonaisilla lehtien nimillä. Lyhenteistä ja niihin vertautuvista kokonaisista nimistä on olemassa valmis taulukko.

2. Stringissä olevat julkaisunumerot tulee muuntaa päivämääriksi. Julkaisunumerot ja niitä vastaavat päivämäärät ovat valmiiksi olemassa erillisessä taulukossa.

3. Stringissä voi olla useampia ilmestymisnumeroita ja lehtiarvoja, esim.

   KÄYTETTY AIKAISEMMIN (lehtiarvo1) (ilmestymisnumero1-1), (ilmestymisnumero1-2), (ilmestymisnumero1-3), (lehtiarvo2) (ilmestymisnumero2-1)

   Nämä tulee eritellä omiin nodeihinsa seuraavasti

```
<publishinghistory>
             <publishing>
             <pubdate> ilmestymisnumero1-1</pubdate>
             <publication> lehtiarvo1</publication>
             </publishing>
             <publishing>
             <pubdate> ilmestymisnumero1-2</pubdate>
             <publication> lehtiarvo1</publication>
             </publishing>
             <publishing>
             <pubdate> ilmestymisnumero1-3</pubdate>
             <publication> lehtiarvo1</publication>
             </publishing>
             <publishing>
             <pubdate> ilmestymisnumero2-1</pubdate>
             <publication> lehtiarvo2</publication>
             </publishing>
</publishinghistory>
```

   (ylläolevan esimerkin lehtiarvot ja ilmestymisnumerot täytyy tietysti ensin kovertoida kokonaisiksi nimiksi ja päivämääriksi)

4. Stringin tunnistamiseen on suunniteltu algoritmi jota voi käyttää alkuun pohjana.

5. Uuden publishinghistory-noden sijainti on historyinfo-noden jälkeen.